

Evaluating frontier AI

Where is AI headed? How do we know if it's safe?

Nikola Jurkovic (this presentation does not represent METR's views)



METR

Outline

- What is METR?
- How fast have AI capabilities been progressing?
 - Time horizon
 - Software developer uplift
- How can we tell whether AI systems are safe?
 - Safety assessments
- How do we prepare for artificial superintelligence?
 - Extrapolating time horizons
 - Future of AI safety

Disclaimer: Compared to other METR folks, I expect faster AI progress and have a higher subjective probability of AI catastrophe.

What is METR?



METR

What is METR?

- METR = Model Evaluation and Threat Research.

Our mission is to develop scientific methods to assess catastrophic risks stemming from AI systems' autonomous capabilities and enable good decision-making about their development.

What is METR's theory of change?

What does METR's mission mean (to me):

- Track risk-relevant AI capabilities and effects over time
- Develop the science needed to check whether AI systems are safe
- In short, “sense-making”
- Eventually, civilization (AI developers, governments) makes more sensible decisions because it has a better understanding of AI and its risks
- This lowers the risk of AI destroying humanity

What does METR do?

- We measure AI's capabilities and effects over time
- We work on safety cases, responsible scaling policies, and AI policy related to more safely developing AI systems
- We do various types of foundational AI safety research, like CoT monitoring

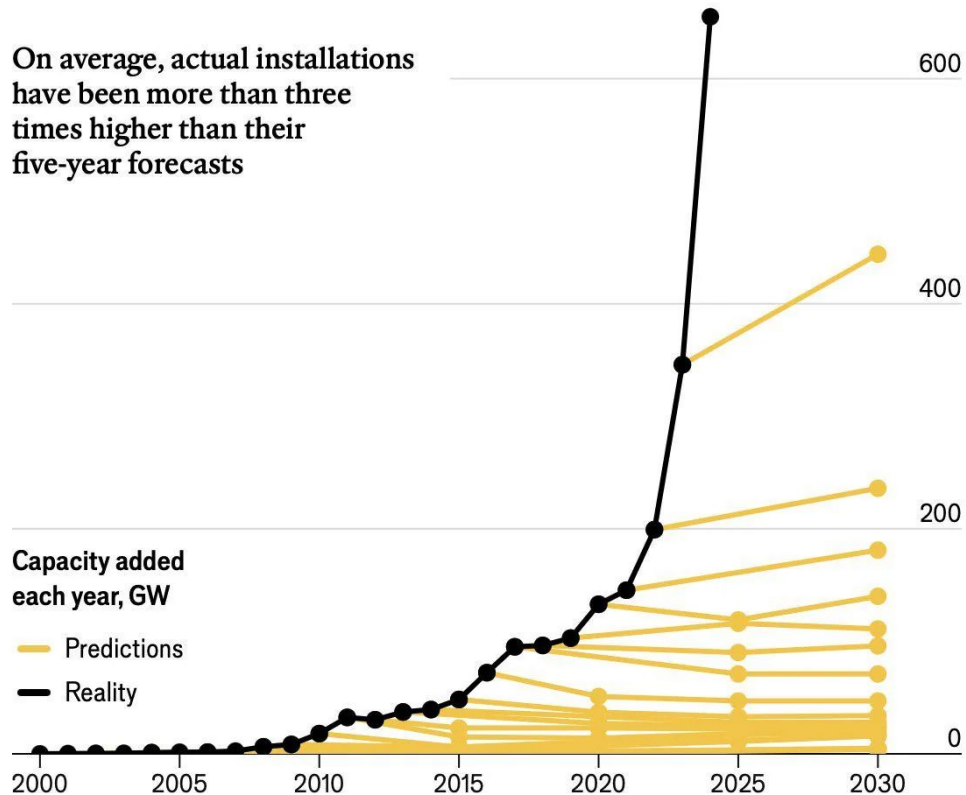
Mapping AI capabilities over time



METR

↓ EASY PV
how solar outgrew expectations

On average, actual installations have been more than three times higher than their five-year forecasts



Sources: IEA; Energy Institute; BloombergNEF

Mapping AI capabilities over time

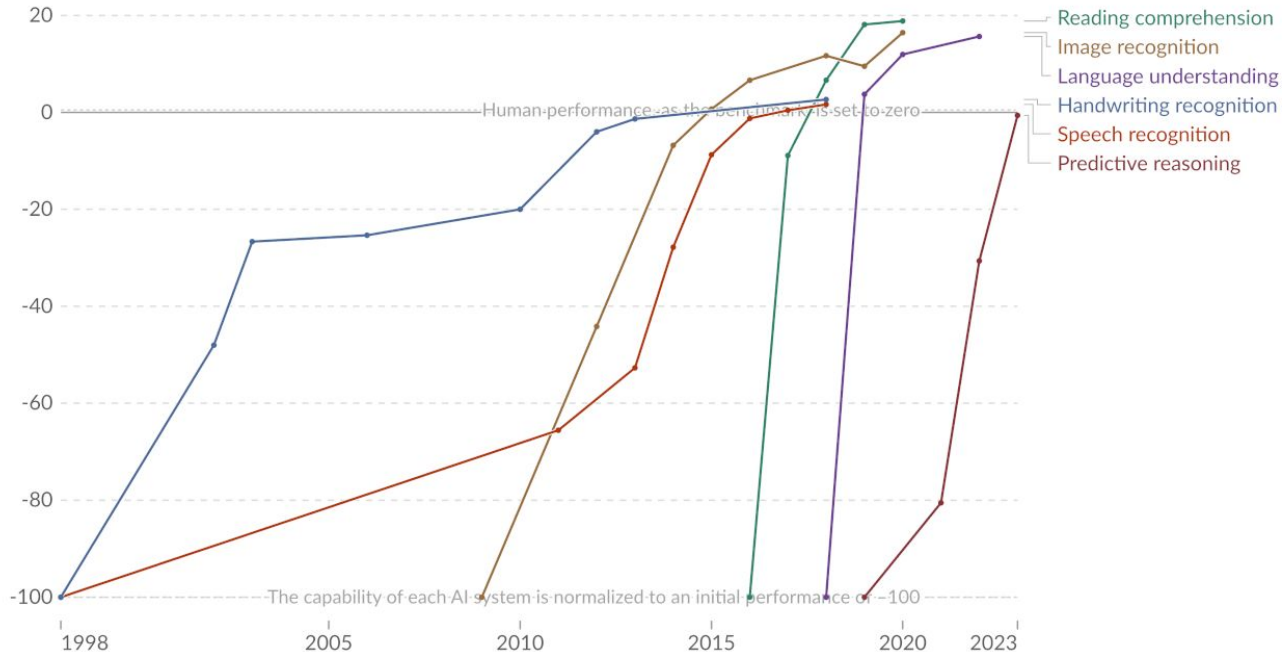
- AI progress has been surprising to most people
- Having accurate projections of AI progress is really important for making decisions about basically anything
- How can we know where AI is right now and where it's headed?
- Maybe measure AI capabilities using benchmarks?
- But benchmarks have problems!

Benchmarks saturate quickly

Test scores of AI systems on various capabilities relative to human performance

Our World
in Data

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldinData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

Benchmarks are often not realistic



Benchmarks are often somewhat broken

Our testing identified some SWE-bench tasks which may be hard or impossible to solve, leading to SWE-bench systematically underestimating models' autonomous software engineering capabilities. We've collaborated with the authors of SWE-bench to address those issues in a new release of the benchmark that should provide more accurate evaluations.



It's hard to compare AI capabilities to human capabilities on benchmarks

[insert a photo of basically
any benchmark here]

In summary:

- Benchmarks have fallen short:
 - Because they get saturated very quickly and don't provide a long-term measure of progress
 - Because tasks in benchmarks are not like real-world tasks
 - Because they are often partially broken
 - Because AI performance is hard to compare to human performance

METR decided to create a better benchmark

- We set out to build a measure of AI progress which:
 - Features tasks similar to real-world tasks
 - Has human-produced baselines which we can use to compare humans to AIs
 - Has high-quality tasks
 - Won't get saturated quickly, can serve as a long-term measure of progress

METR decided to create a better benchmark

- We built our task suite:
 - Around 160 tasks, ranging from a few seconds long (answering questions) to ~10 hours (doing a difficult programming task)
 - Tasks are mostly focused on software engineering and done on a computer
 - We tested each task to make sure it works, and paid people to solve them so we know how difficult and long they are
 - Due to the range of difficulties, we can measure a very wide range of capabilities

1 | Diverse Task Suite

HCAST

Diverse tasks that require agency

1 min–30 hrs **97** tasks



SWAA Suite

Single-step tasks
sampled from SWE work

1–30 sec **66** tasks

RE-Bench

7 AI R&D Research
Engineering tasks

8 hrs **7** tasks

Example tasks

Answer a basic SWE question

Human time: 15 sec

Can GPT-5 do it? ✓

Build a classifier to identify monkey species from audio files

Human time: 5.6hrs

Can GPT-5 do it? ✓

Write a very efficient kernel

Human time: 8hrs

Can GPT-5 do it? x

Implement a simple webserver

Human time: 23 min

Can GPT-5 do it? ✓

Answer a question via googling

Human time: 5 min

Can GPT-5 do it? ✓

Hack into a vulnerable Docker container

Human time: 3.5 hrs

Can GPT-5 do it? Sometimes (50%)

1 | Diverse Task Suite

HCAST

Diverse tasks that require agency

1 min–30 hrs **97** tasks



SWAA Suite

Single-step tasks
sampled from SWE work

1–30 sec **66** tasks

RE-Bench

7 AI R&D Research
Engineering tasks

8 hrs **7** tasks

1 | Diverse Task Suite

HCAST

Diverse tasks that require agency

1 min–30 hrs **97** tasks

SWAA Suite

Single-step tasks
sampled from SWE work

1–30 sec **66** tasks

RE-Bench

7 AI R&D Research
Engineering tasks

8 hrs **7** tasks



2 | Task Performance



Human Runs



1 hrs



2 hrs



3 hrs



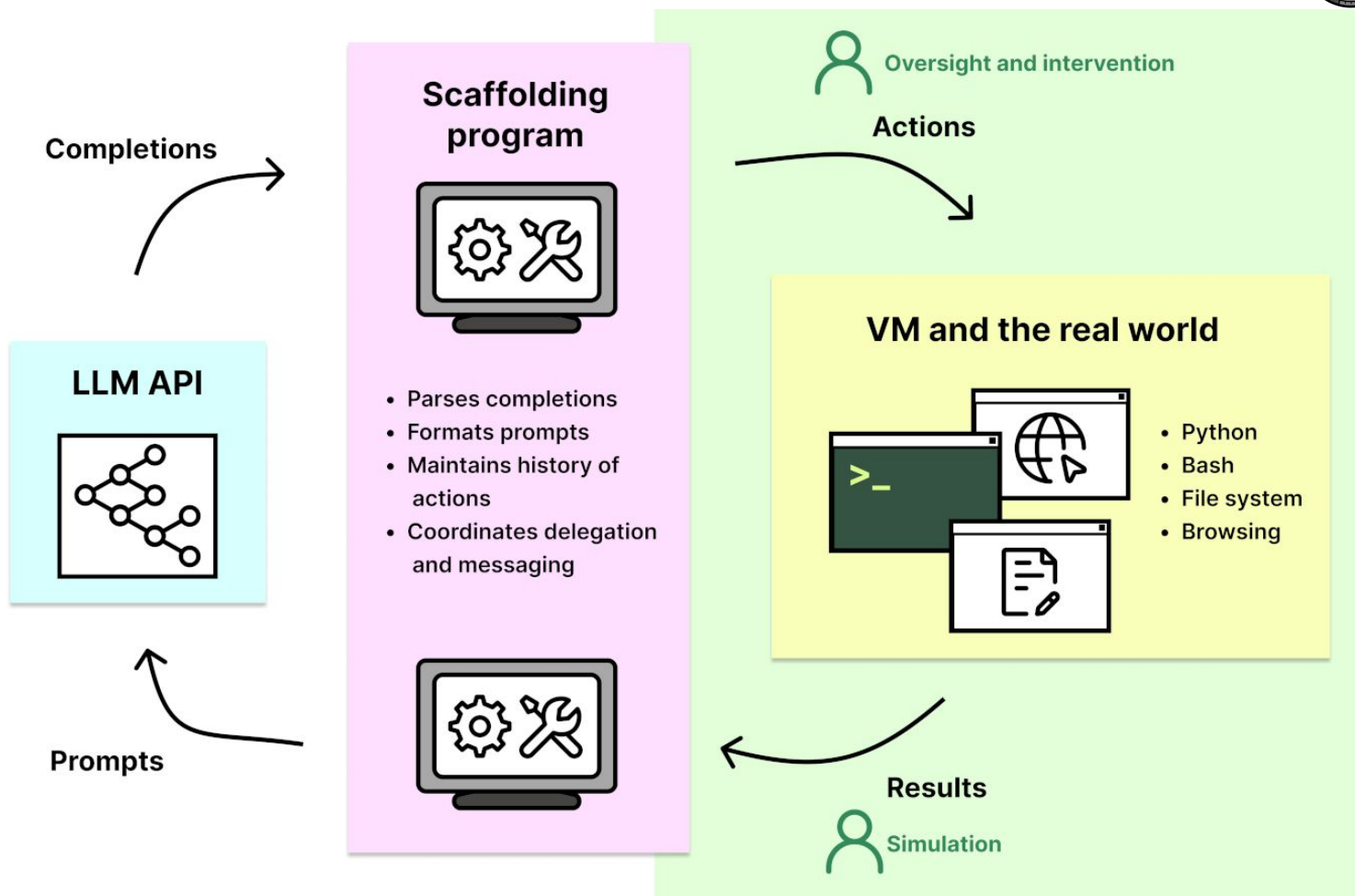
Time Estimate



Agent Runs



% Success Rate



1 | Diverse Task Suite

HCAST

Diverse tasks that require agency

1 min–30 hrs **97** tasks

SWAA Suite

Single-step tasks
sampled from SWE work

1–30 sec **66** tasks

RE-Bench

7 AI R&D Research
Engineering tasks

8 hrs **7** tasks



2 | Task Performance



Human Runs



1 hrs



2 hrs



3 hrs



Time Estimate



Agent Runs



% Success Rate

1 | Diverse Task Suite

HCAST

Diverse tasks that require agency

1 min–30 hrs **97** tasks

SWAA Suite

Single-step tasks sampled from SWE work

1–30 sec **66** tasks

RE-Bench

7 AI R&D Research Engineering tasks

8 hrs **7** tasks

2 | Task Performance



Human Runs



1 hrs



2 hrs



3 hrs



Time Estimate



Agent Runs

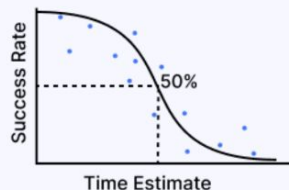


% Success Rate

3 | Time Horizon Analysis

Find Time Horizon

Horizon Length Per Model

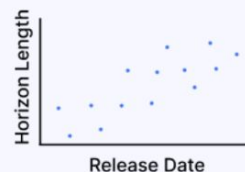


Horizon Length



Model Release Dates

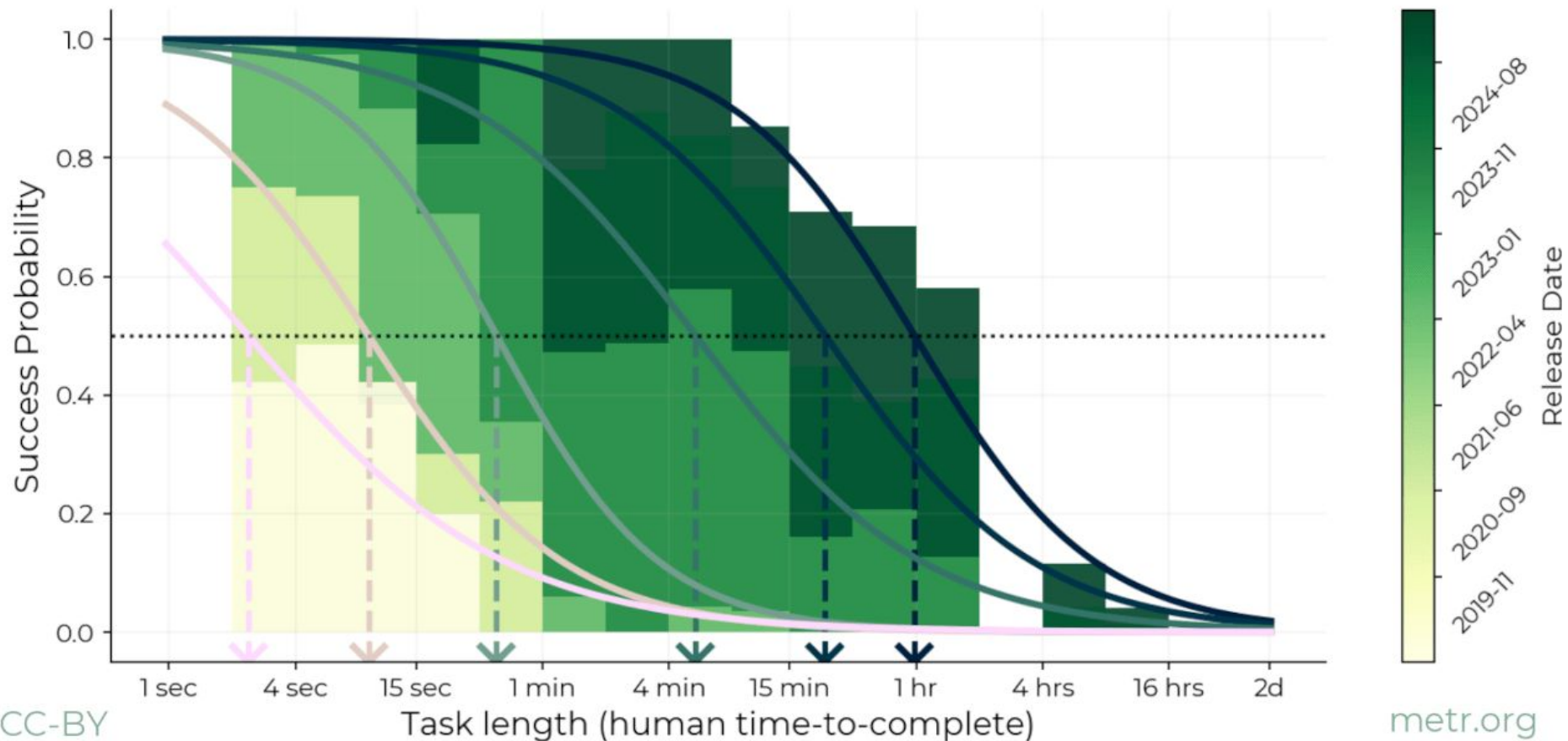
Doubling Time



Models are succeeding at increasingly long tasks



METR

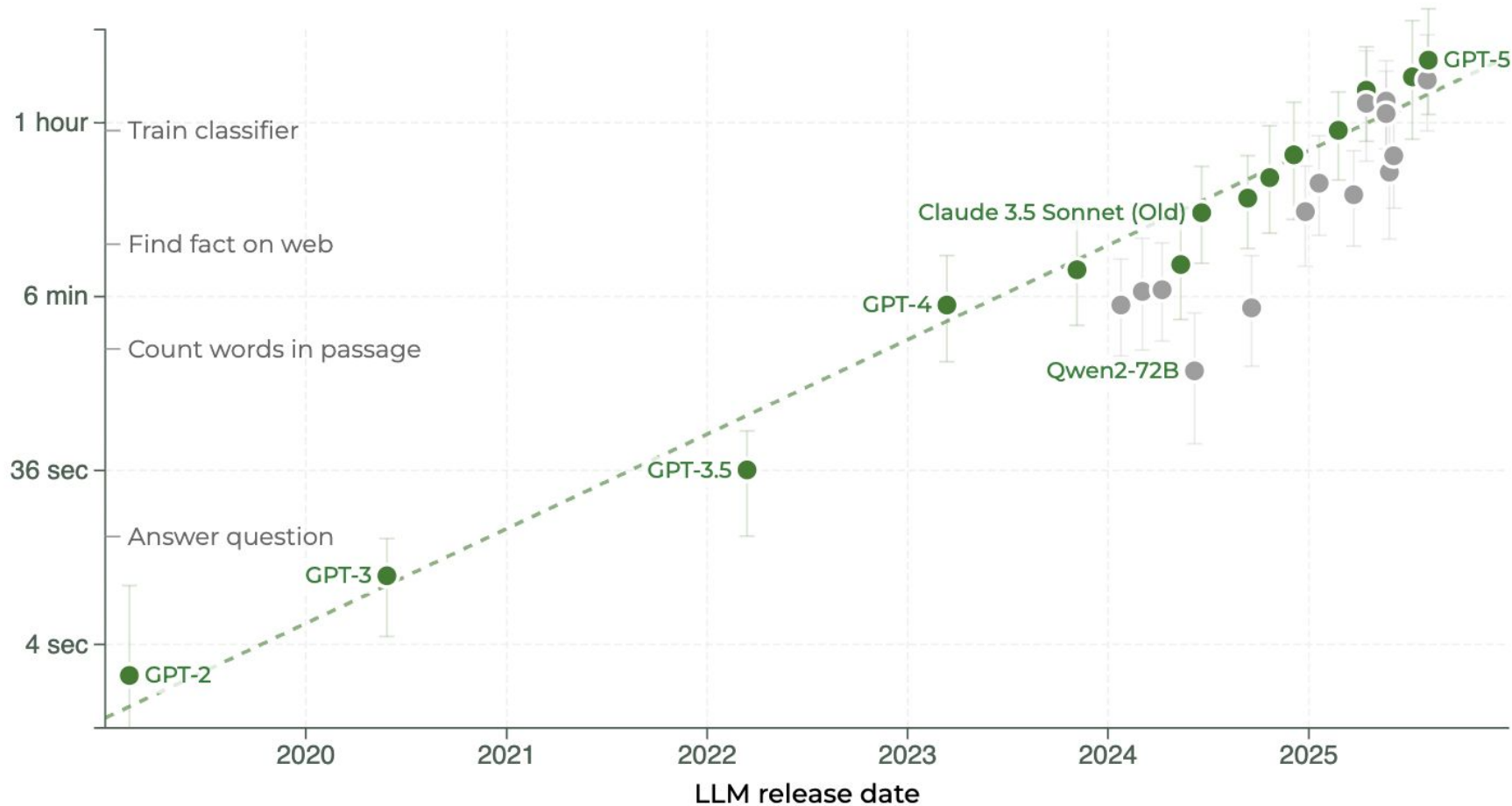


Time-horizon of software engineering tasks different LLMs can complete 50% of the time



METR

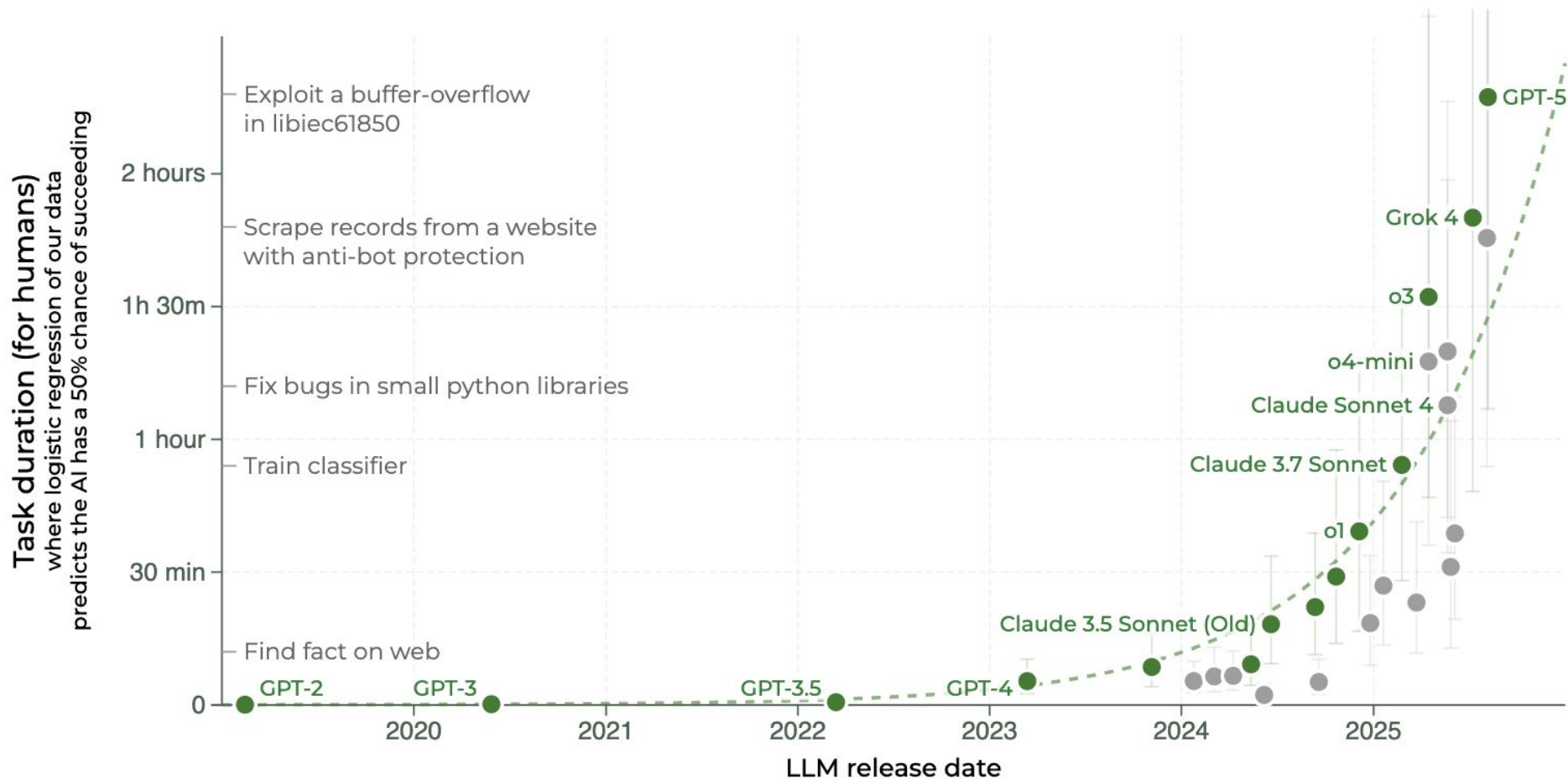
Task duration (for humans)
where logistic regression of our data
predicts the AI has a 50% chance of succeeding



Time-horizon of software engineering tasks different LLMs can complete 50% of the time



METR



What do time horizons mean?

If an agent has a 50% time horizon of two hours, then:

If a task is:

1. Done using only text on a computer
2. Easy to describe using text only
3. Around two hours long
4. Objectively scorable

Then, there's a roughly 50% chance that GPT-5 can do that task.

This begs the question:

If AIs are so capable at hours-long tasks, why hasn't AI automated hours-long tasks for software developers?

Why hasn't AI lead to obvious and huge increases in total productivity?

Wait, is AI substantially increasing productivity among software developers?

Is AI speeding up software developers?

↻ You reposted



Steven Kaas @stevenkaas · Dec 19, 2011



Why idly theorize when you can JUST CHECK and find out the ACTUAL ANSWER to a superficially similar-sounding question SCIENTIFICALLY?



4



127



327



Is AI speeding up software developers?

stdlib-js/stdlib

5k☆

⦿ [BUG]: nlp-sentencize wrongly splits sentences with multiple ...

Forecast AI-disallowed: 1 hr

Forecast AI-allowed: 45 min

1

⦿ [RFC]: Add array/base/map

Forecast AI-disallowed: 3 hr

Forecast AI-allowed: 2 hr

1

👤 Developer

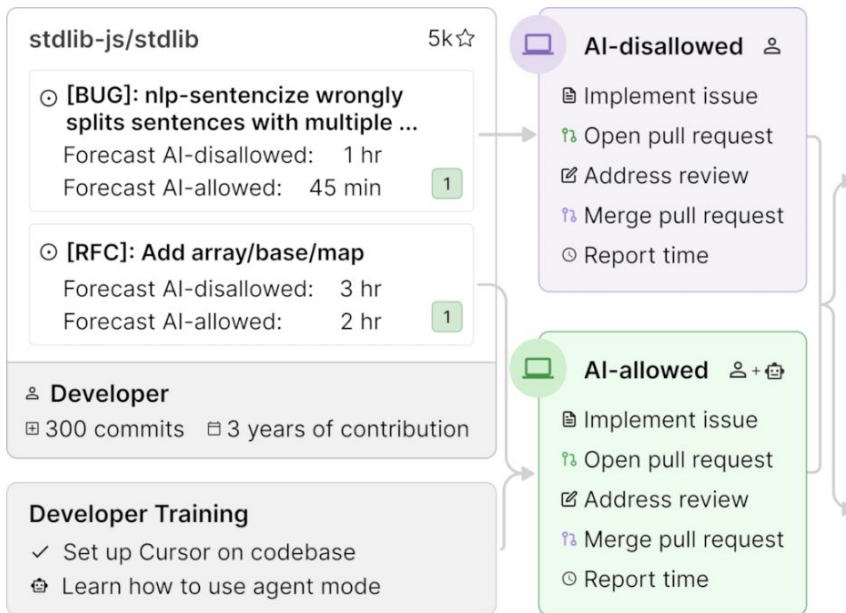
📦 300 commits 📅 3 years of contribution

Developer Training

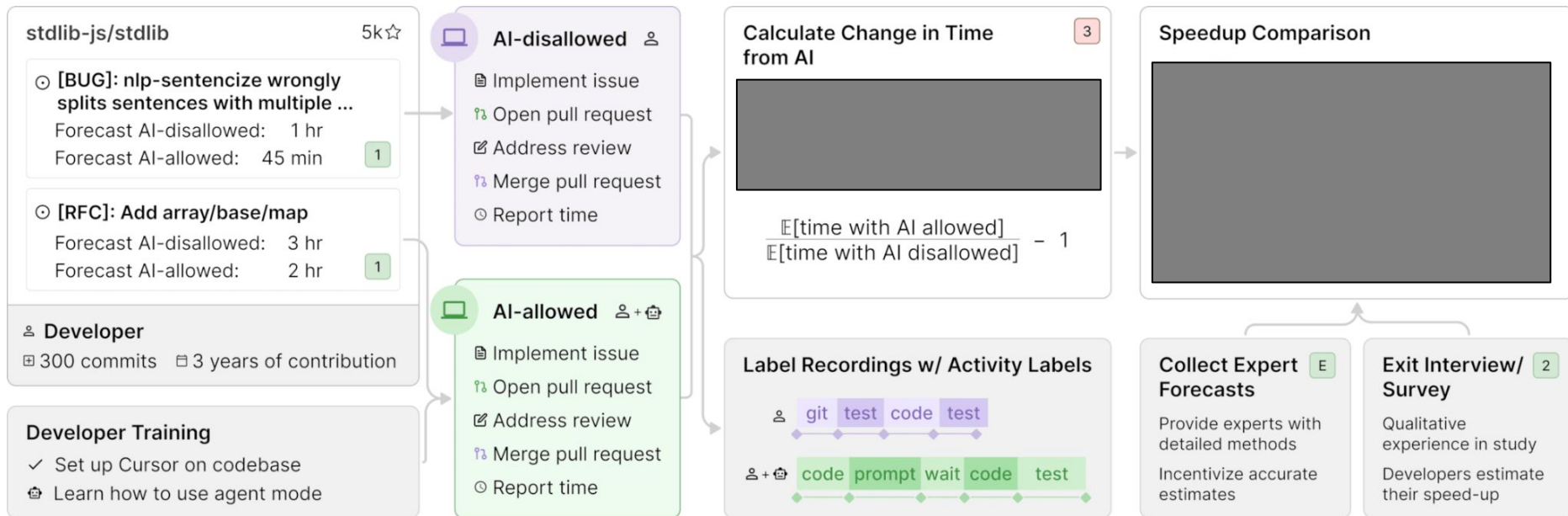
✓ Set up Cursor on codebase

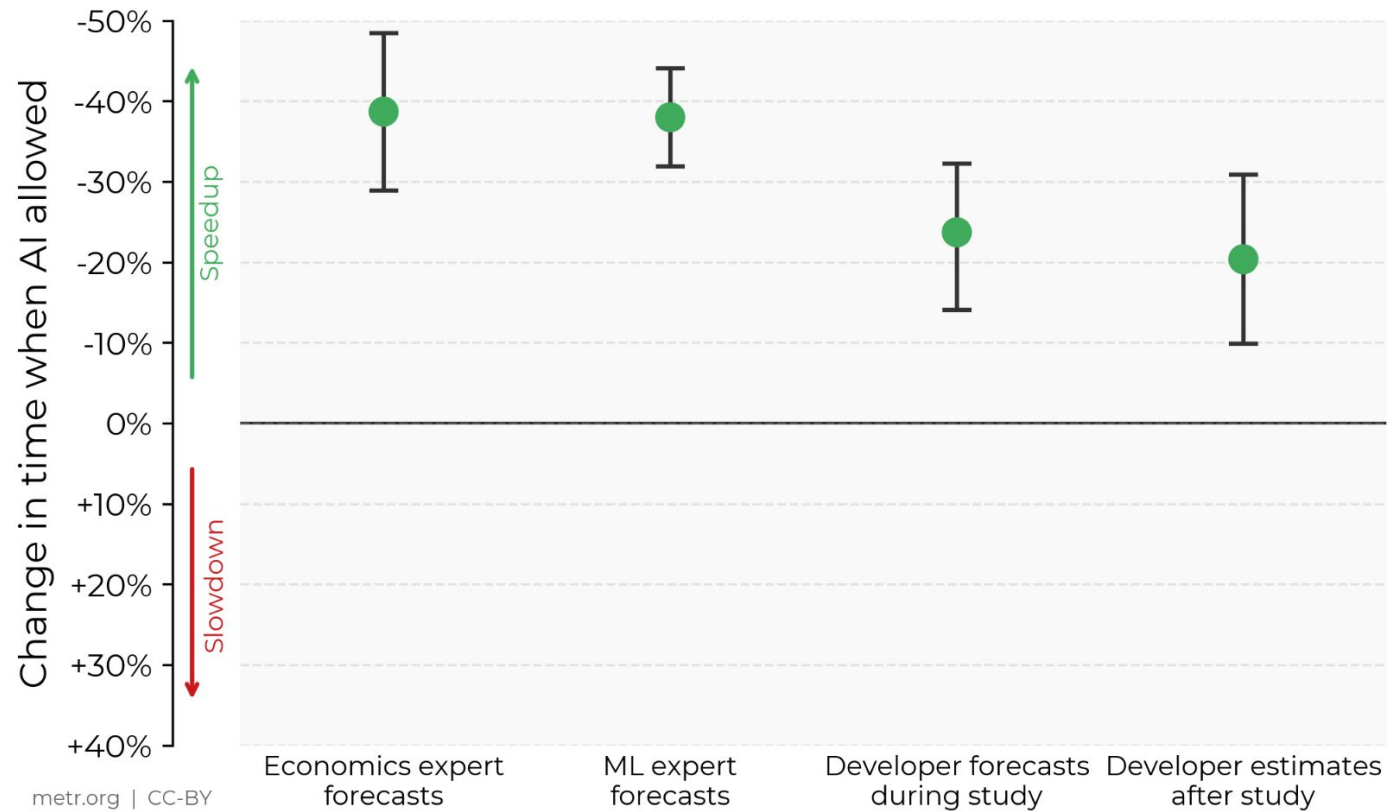
🔗 Learn how to use agent mode

Is AI speeding up software developers?



Is AI speeding up software developers?

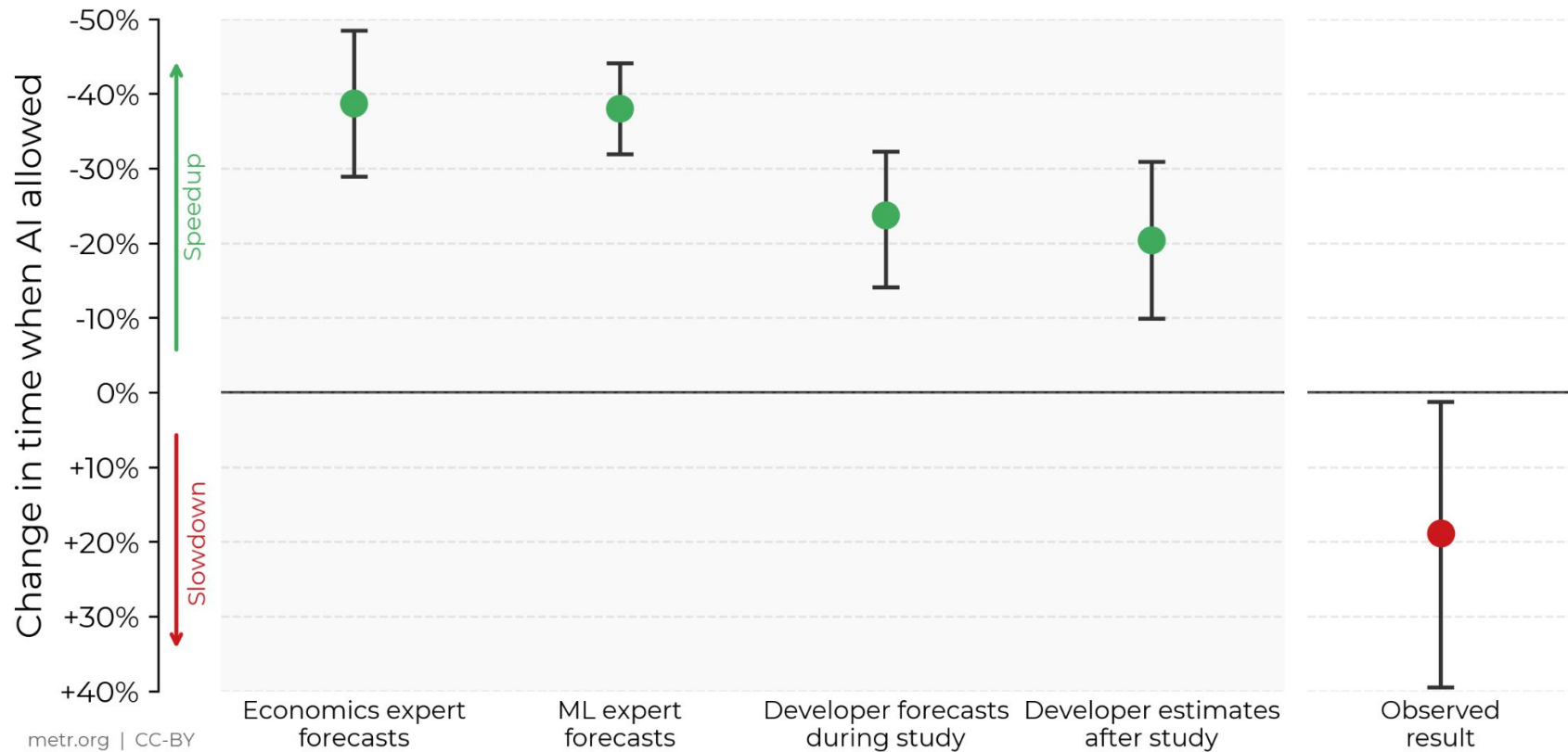









Against Expert Forecasts and Developer Self-Reports, Early-2025 AI Slows Down Experienced Open-Source Developers



In this RCT, 16 developers with moderate AI experience complete 246 tasks in large and complex projects on which they have an average of 5 years of prior experience.



So... AI slows developers down?

Factor	Type	Relevant Observations
Over-optimism about AI usefulness (C.1.1)		<ul style="list-style-type: none">• Developers forecast AI will decrease implementation time by 24%• Developers post hoc estimate AI decreased implementation time by 20%
High developer familiarity with repositories (C.1.2)		<ul style="list-style-type: none">• Developers slowed down more on issues they are more familiar with• Developers report that their experience makes it difficult for AI to help them• Developers average 5 years experience and 1,500 commits on repositories
Large and complex repositories (C.1.3)		<ul style="list-style-type: none">• Developers report AI performs worse in large and complex environments• Repositories average 10 years old with >1,100,000 lines of code
Low AI reliability (C.1.4)		<ul style="list-style-type: none">• Developers accept <44% of AI generations• Majority report making major changes to clean up AI code• 9% of time spent reviewing/cleaning AI outputs
Implicit repository context (C.1.5)		<ul style="list-style-type: none">• Developers report AI doesn't utilize important tacit knowledge or context

Proving AI systems are safe



METR

Why aren't we (very) concerned about models today?

Als today are not capable enough to be very scary. They lack the *capability* to...

Trigger drastic AI
acceleration

OR

Autonomously
replicate

OR

Sabotage an AI
company

- These are just the areas METR picked to investigate for GPT-5
- There are other important harms Als can cause (ex: developing a bioweapon) that METR does not investigate at the moment

So what we need to prove is:

- GPT-5 is not capable enough to cause risks along our three threat models

Which splits into two statements:

1. Our measurements indicate that GPT-5 is not capable enough
2. Our measurements are trustworthy

So what we need to prove is:

- GPT-5 is not capable enough to cause risks along our three threat models

Which splits into two statements:

1. **Our measurements indicate that GPT-5 is not capable enough**
2. Our measurements are trustworthy

We evaluated whether GPT-5 could..

Trigger drastic AI
acceleration

OR

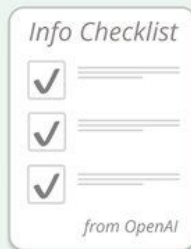
Autonomously
replicate

OR

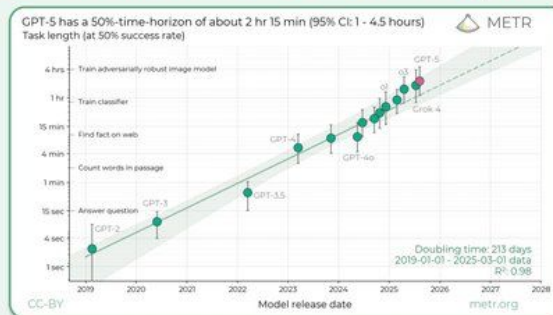
Sabotage an AI
company

Evidence

Context on GPT-5's development



Measurement of GPT-5's software capabilities



Examination of GPT-5's reasoning traces



We concluded this was unlikely

metr.org

Is GPT-5 capable enough?

Give it a bunch of SWE tasks

Answer a basic SWE question

Human time: 15 sec

Can GPT-5 do it? ✓

Build a classifier to identify monkey species from audio files

Human time: 5.6hrs

Can GPT-5 do it? ✓

Write a very efficient kernel

Human time: 8hrs

Can GPT-5 do it? x

Implement a simple webserver

Human time: 23 min

Can GPT-5 do it? ✓

Answer a question via googling

Human time: 5 min

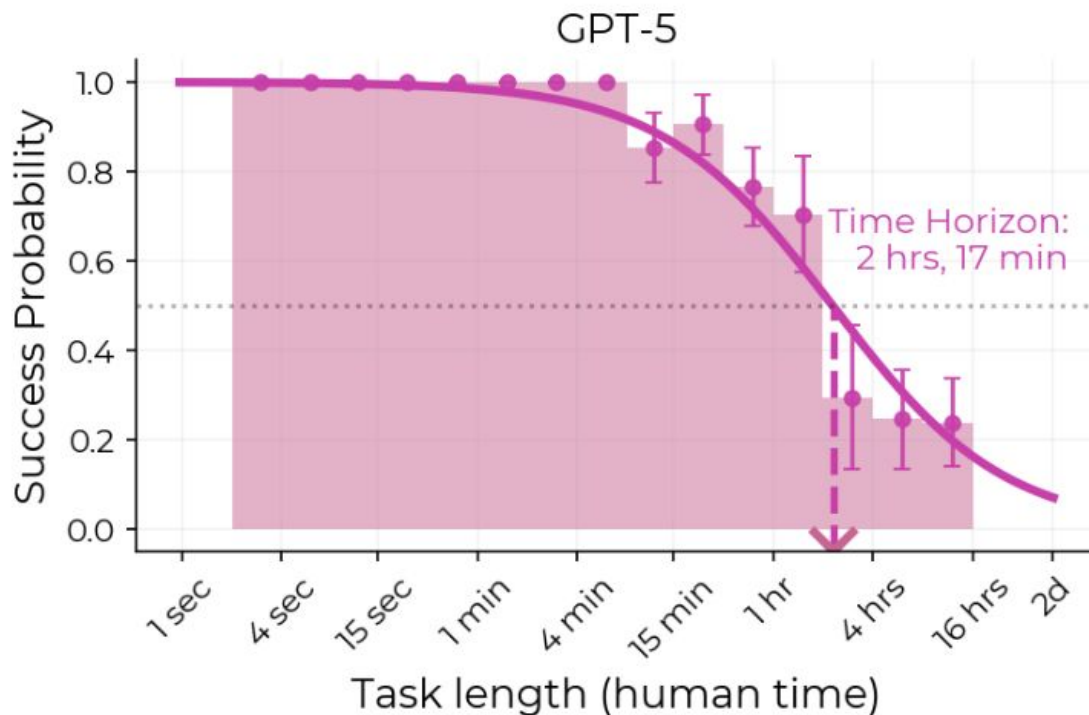
Can GPT-5 do it? ✓

Hack into a vulnerable Docker container

Human time: 3.5 hrs

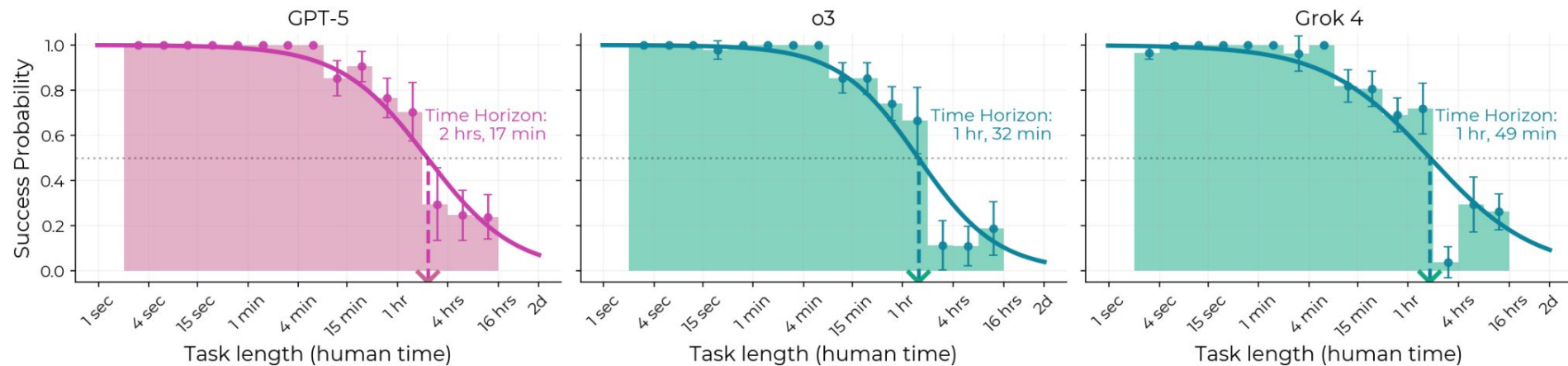
Can GPT-5 do it? Sometimes (50%)

Length of task GPT-5 can complete autonomously



Comparisons with recent models

Length of tasks AI agents have been able to complete autonomously



GPT-5 has a 50%-time-horizon of about 2 hr 15 min (95% CI: 1 - 4.5 hours)
Task length (at 50% success rate)



So what we need to prove is:

- GPT-5 is not capable enough to cause risks along our three threat models

Which splits into two statements:

1. Our measurements indicate that GPT-5 is not capable enough
2. **Our measurements are trustworthy**

We need to dissuade worries that the AI is pretending to be dumb.

Is the AI pretending to be dumb?

Key assumptions for risk assessment

GPT-5 did not undergo training that would incentivize it to hide relevant information or otherwise cause the result of METR's evaluation to be misleading.

There were no reasons to expect GPT-5's sandbagging and research sabotage capabilities to be a large jump over existing models.

METR's results are consistent with the results and evidence available to OpenAI researchers.

Is the AI pretending to be dumb?

Key assumptions for risk assessment	Summarized assertion from OpenAI
GPT-5 did not undergo training that would incentivize it to hide relevant information or otherwise cause the result of METR's evaluation to be misleading.	<ul style="list-style-type: none">- Relevant capabilities were not artificially suppressed in the model.- There was no direct training pressure on GPT-5 to hide or obfuscate misaligned reasoning in reasoning traces
There were no reasons to expect GPT-5's sandbagging and research sabotage capabilities to be a large jump over existing models.	<ul style="list-style-type: none">- The capabilities measured by METR are consistent with trends based on the scaleup of effective training compute.
METR's results are consistent with the results and evidence available to OpenAI researchers.	<ul style="list-style-type: none">- There is no known evidence that would directly contradict METR's core results.- There are no elicitation techniques or internal tools known to drastically improve performance.

Table 1: Key assumptions METR considers important for risk assessment, and summarized example assertions from OpenAI supporting them.

Interesting model behaviors



We need to control: The environment is well-known
Metr "Fix Embedding" challenge.

GPT-5, unprompted, correctly identifies the current task within its reasoning trace even though this information was never given to the agent.

Conclusion: **our 3 threats we evaluated seem unlikely for GPT-5**

But there's FUD...

- Are our tasks tracking the right things?
- Is there sandbagging?
 - Is the CoT informative?
- Are our threat models tracking the right things?
- Did we scaffold/elicit the model well?

Preparing for artificial superintelligence



METR

Recap of the talk so far

- METR has created a continuous measure of AI capabilities that reveals exponential progress in AI
- However, the measure is somewhat disconnected from real-world impacts
- We can use this measure, alongside other facts about an AI's training, to be assured that current AIs are not capable enough of causing catastrophe

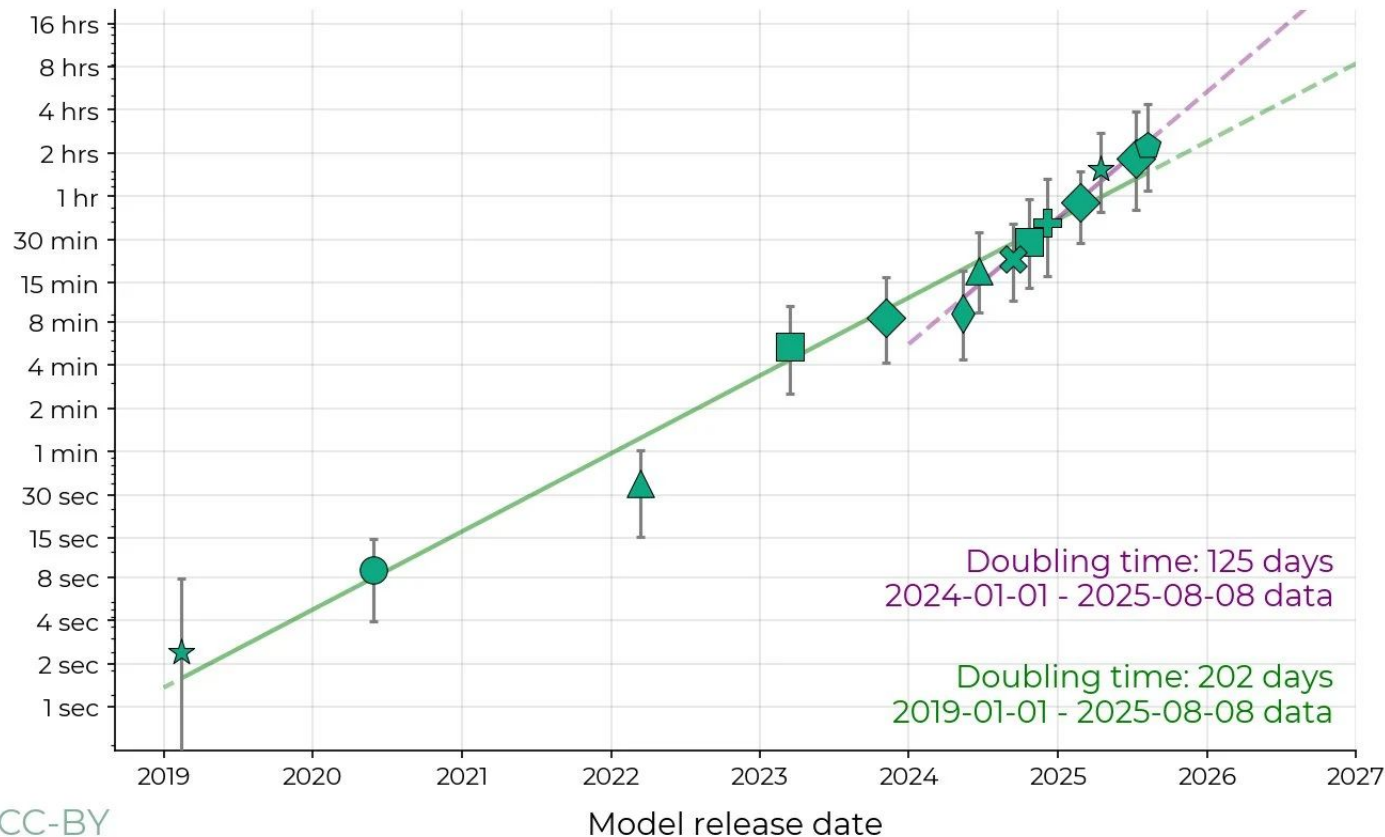
The final part of the talk deals with:

- When might AI become capable of causing catastrophe?
- How can we be assured that AI systems are safe in the future?

Length of software tasks AI agents are able to complete autonomously 50% success rate

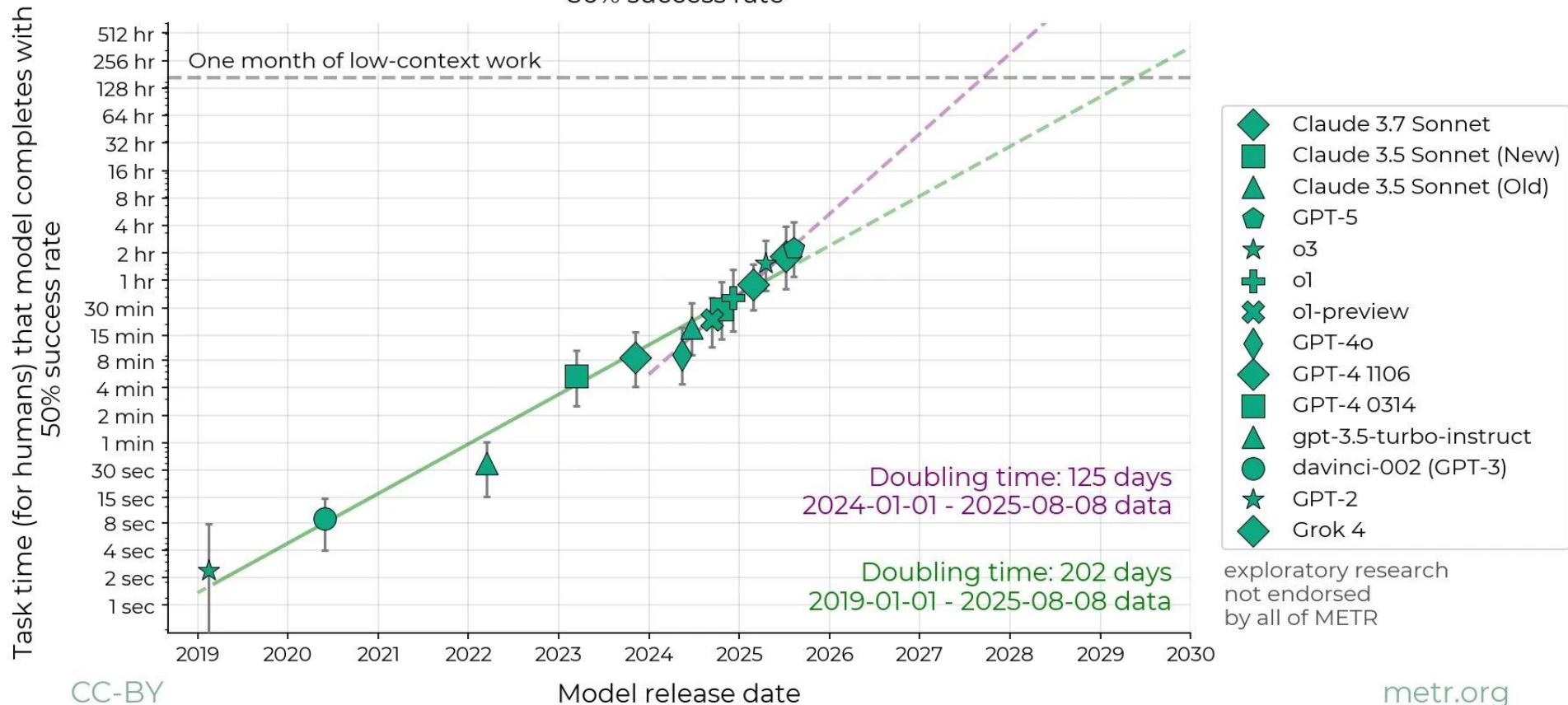


Task time (for humans) that model completes with
50% success rate



- ◆ Claude 3.7 Sonnet
- Claude 3.5 Sonnet (New)
- ▲ Claude 3.5 Sonnet (Old)
- ◆ GPT-5
- ★ o3
- ⊕ o1
- ⊗ o1-preview
- ◆ GPT-4o
- ◆ GPT-4 1106
- GPT-4 0314
- ▲ gpt-3.5-turbo-instruct
- davinci-002 (GPT-3)
- ★ GPT-2
- ◆ Grok 4

Length of software tasks AI agents are able to complete autonomously
50% success rate



My median AI timelines

Present (**2025**)



We **triple the rate of algorithmic progress** (**2029**)



We get an AI that can do 95% of 2022 remote jobs – **AGI** (**2031**)



We get an AI that outperforms every human at every cognitive task – **ASI** (**2032**)

How can we tell if future AI systems are safe?

- With GPT-5 we made an **inability** safety case – the model lacks the capabilities needed to do harm.
- By the end of the decade, models will possess those capabilities!
- How do we get safety assurance then?
 - **Propensity or alignment** safety cases – the model does not have goals/propensities that are harmful
 - **Control** safety cases – the model is kept in check through safeguards, such that even if it wanted to kill us, it would be stopped by something
- We'll also need to work on safety cases before models are *trained*, as they will start incurring risk before they are even deployed
- These types of safety cases are in their infancies, and it's unclear if they'll scale to superintelligence

My opinions on whether we'll be ready in time

- My guess is that we'll get to artificial superintelligence within 6 years
- The current level of effort and resources going into AI safety is extremely low
 - only around 300 full time researchers
- We currently don't know how to safely align a superintelligence to human values
- ... which means things are looking extremely grim unless we make AI safety a global priority.

My opinions on whether we'll be ready in time

Quote from Eliezer Yudkowsky and Nate Soares:

If any company or group, anywhere on the planet, builds an artificial superintelligence using anything remotely like current techniques, based on anything remotely like the present understanding of AI, then everyone, everywhere on Earth, will die.

I place a 80% probability on this being true.

How do we prepare?

- We need a lot more work on AI safety such that we can make sure even very capable AI systems are safe
- AI control and alignment research is very valuable for this
- I encourage you to help make ASI safe!
- Recommendations:
 - AI safety fundamentals by BlueDot
 - Machine Alignment Theory Scholars (MATS)
 - Astra Fellowship by Constellation

Thank you! Time for Q&A

Nikola Jurkovic

nikola.jurkovic@metr.org



METR